DOCUMENT RESUME

ED 273 658 TM 860 507

AUTHOR Becker, Betsy Jane
TITLE Asymptotic Distributions for Tests of Combined Significance.
PUB DATE Apr 86
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Correlation; Effect Size; *Hypothesis Testing; Mathematical Models; *Sampling; Simulation; *Statistical Distributions; *Statistical Significance; Statistical Studies; Teacher Attitudes; *Test Theory
IDENTIFIERS *Asymptotic Distribution; Convergence Theory; *Power (Statistics)

ABSTRACT

This paper discusses distribution theory and power computations for four common "tests of combined significance." These tests are calculated using one-sided sample probabilities or p values from independent studies (or hypothesis tests), and provide an overall significance level for the series of results. Noncentral asymptotic sampling distributions are derived for four combined significance tests (the method of "adding z's," Fisher's method of "adding logs," George's logit summary, and Edgington's mean-p method), and a simulation study is used to examine the accuracy of the distributions in finite samples. The paper also presents a noncentral asymptotic sampling distribution for z(p), the standard normal deviate associated with the one-sided p value, which is used to derive the distributions of statistics used in tests of combined significance. A data set from research on teaching is used to illustrate estimation of power of the combined significance tests to detect several plausible patterns of study outcomes. Recommendations are made for the use of the tests of combined significance. Appendix A presents notation for the results of studies using two-sample comparisons, and outlines how the distributions would be applied to such results. Appendix B provides a proof of the convergence theorem. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED273658

Asymptotic Distributions for Tests of Combined Significance

Betsy Jane Becker

Michigan State University

April, 1986

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. J. Becker

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the Annual Meeting of the American Educational Research Association

San Francisco, April, 1986.

Running head: ASYMPTOTIC DISTRIBUTIONS

TM 860 507

Asymptotic Distributions for Tests of Combined Significance

Abstract

Tnis paper discusses distribution theory and power computations for four common "tests of combined significance." These tests are calculated using one-sided sample probabilities or p values from independent studies (or hypotnesis tests), and provide an overall significance level for the series of results. Noncentral asymptotic sampling distributions are derived for four combined significance tests (the method of "adding z's," Fisher's method of "adding logs," George's logit summary, and Edgington's mean-p method), and a simulation study is used to examine the accuracy of the distributions in finite samples. The paper also presents a noncentral asymptotic sampling distribution for z(p), the standard normal deviate associated with the one-sided p value, which is used to derive the distributions of statistics used in tests of combined significance. A data set from research on teaching is used to illustrate estimation of power of the combined significance tests to detect several plausible patterns of study outcomes. Finally, recommendations are made for the use of the tests of combined significance.

Asymptotic Distributions for Tests of Combined Significance

In recent years meta-analysis (Glass, 1976), or the quantitative synthesis of research findings, has become an increasingly popular method for reviewing empirical research literature in education and psychology. Research reviewers often wish to assess the size or strength of effects found in a particular research domain. For example, White (1982) used quantitative methods to synthesize the literature on the strength of the relationship between socioeconomic status and achievement. Thus one approach to meta-analysis involves accumulating and analyzing quantitative measures of effect magnitude from series of studies.

Another approach to quantitative reviewing involves the calculation of an "overall level of significance" (Rosenthal, 1978, p. 185) or probability for the series of studies. This overall probability is based on a "test of combined significance," a summary calculated using the sample significance values from the individual studies. Though statisticians have used combined significance methods since the early 1930's (Fisher, 1932; Tippett, 1931), social scientists have shown only passing interest (e.g., Jones & Fiske, 1953; Mosteller & Bush, 1954) until recently. Rosenthal (1978) brought the approach to the attention of social scientists with his important review of eight combined significance tests, and several contemporaneous papers (Cooper, 1979; Cooper & Rosenthal, 1980) helped promote the approach through its application.

The "combined significance" approach does not provide direct information about the strength of research findings because sample significance values depend on the sizes of the samples in the individual studies as well as the magnitudes of the effects. However, the approach is broadly applicable because it requires few assumptions, and because sample significance values are often available in study reports even when effect-

magnitude measures are not.

This paper presents noncentral asymptotic distributions for four of the most widely used combined significance tests: the method of "adding z's" proposed by Stouffer and his colleagues (1949), Fisher's (1932) method of "adding logs" (i.e., logarithms of the sample probabilities), George's (1977) logit summary, and Edgington's (1972b) mean-p method. These tests are examined primarily because they are frequently applied in substantive reviews, but also because they satisfy certain mathematical requirements which are discussed below.

The distribution theory given here provides reviewers with a reasonably simple means of estimating power for these tests, thus providing an overall indication of the power of a synthesis of significance values. Furthermore, approximate power values computed for the tests can be compared to evaluate whether one test would be most likely to detect interesting study outcomes, and thus should be employed in a review.

Section one presents notation for the results of a series of studies, each of which involves a hypothesis test based on the t distribution. Two measures of study outcomes are described (the effect-size estimate and the sample significance value). The following sections present the null and alternative hypotheses for tests of combined significance, the formulas for the combined significance tests, and a brief discussion of the power of tests of combined significance. Section five begins with a presentation of a convergence theorem which enables the derivation of the asymptotic distributions. Next the noncentral asymptotic distribution of z(p), the standard normal deviate for the one-sided p value, is stated (Lambert, 1978), the noncentral asymptotic distributions for the combined significance tests are presented, and their accuracy is examined. In section six the asymptotic distributions are used to estimate power for summaries of a set

of studies from Gage's (1978) book The scientific basis of the art of teaching. The last section contains recommendations for the use of the tests and comments about combined significance tests not examined in this paper.

Notation for the Results of a Series of Studies

In order to simplify computations, I study the tests of combined significance in the situation in which each study examines a single parameter via a one-sample t test. A common example is the situation in which each study tests for differences in the means of paired data. Another example is the summary of correlational studies (in which each study tests whether the population correlation ρ is equal to zero).

Consider a series of k studies each examining a hypothesis about a population mean μ_i. Let Y_{ij} be the score of the jth person from the ith study. Denote the ith sample size as n_i and assume that in study i, the Y_{ij} are independently normally distributed with mean μ_i and variance σ_i^2. Thus

$$Y_{ij} \sim N(\mu_i, \sigma_i^2), \quad \text{for } j = 1,\ldots,n_i, \; i = 1,\ldots,k.$$

The parameter under study can be considered as μ_i or as $\delta_i = \mu_i/\sigma_i$ with no loss of generality. This parameter δ_i is the one-sample analogue to the two-sample population standardized mean difference which has come to be commonly known as the effect size. The sample estimate of $\delta_i = \mu_i/\sigma_i$ is the value $g_i = \bar{Y}_i/S_i$, where $\bar{Y}_i$ is the ith sample mean and S_i is the ith standard deviation. The k parameters $\delta_1, \ldots, \delta_k$ are the population values about which inferences are made when sample effect sizes are analyzed.

The Significance Value

The relationship between the effect size and the sample significance value (or p value) arises through the relation of both to the t test. The relationship of the estimate g_i to the one-sample t test is obvious, since

in the ith study,

$$t_i = \sqrt{n_i}(\bar{Y}_i - \mu_i)/S_i = \sqrt{n_i}[(\bar{Y}_i/S_i) - (\mu_i/S_i)], \tag{1}$$

which under the null model that $\mu_i = 0$ equals $\sqrt{n_i}\, g_i$. This t statistic has $n_i - 1$ degrees of freedom.

The one-sided (directional) significance value corresponding to the effect measured by g_i is the upper-tail probability from a t test of the null hypothesis that the ith mean is zero. For the ith study the null model is

$$H_{0i}: \ \mu_i = 0, \tag{2}$$

versus the alternative that the ith experimental population mean exceeds zero. This directional alternative can be written symbolically as

$$H_{1i}: \ \mu_i > 0.$$

To test whether the ith mean equals a particular value μ_0, we simply reparameterize the model with $\mu' = \mu - \mu_0$ and test whether $\mu_i' = \mu_i - \mu_0 = 0$.

The ith population effect size is simply the ith population mean μ_i standardized with the common standard deviation σ_i. Thus the t test that the ith population mean is positive (H_{1i} shown above) is also a test for a positive population effect size. That is, testing H_{0i} versus H_{1i} is equivalent to testing the null hypothesis that $\delta_i = 0$ versus the alternative that $\delta_i > 0$, where δ_i is defined as μ_i/σ_i.

The sample significance value from the ith study used in the combined significance tests is the upper-tail p value (p_i) associated with the sample value of t given in (1) above. The significance value p_i is formally

defined in terms of the sample $\underline{t}$ statistic.

$$\underline{p}_i = \int_{\underline{t}_i}^{\infty} f(\underline{t}, \underline{m}_i)\, d\underline{t}, \qquad (3)$$

where $f(\underline{t}, \underline{m}_i)$ is the probability density function for the central $\underline{t}$ distribution with $\underline{m}_i = \underline{n}_i - 1$ degrees of freedom.

Hypotheses for Tests of Combined Significance

A single, very simple, combined null hypothesis is examined by the combined significance tests in the effect-size context. The combined null hypothesis is that all of the studies are of populations in which the individual null hypothesis is true. In other words, the model states that the population effect sizes for all of the studies equal zero. We write

$$H_0: \ \delta_1 = \delta_2 = \ldots = \delta_{\underline{k}} = 0 \qquad (4)$$

where $\delta_{\underline{i}}$ is the population effect size from the $\underline{i}$th study.

When the combined significance tests are based on one-sided significance values (i.e., $\underline{p}$ values defined as in (3) above) the model for the alternative hypothesis is

$$H_1: \ \delta_{\underline{i}} \geq 0, \ \text{for } \underline{i} = 1,\ldots,\underline{k}, \text{ and at least one } \delta_{\underline{i}} > 0. \qquad (5)$$

Almost all applications of combined significance tests in the social science literature and elsewhere have used one-sided $\underline{p}$ values. Consequently, the alternative shown in (5) is the one typically tested in meta-analysis. When the null model given in (4) is rejected on the basis of a summary of one-sided significance values, the reviewer can conclude that <u>at least one</u> population effect size is nonzero and positive.

The Combined Significance Tests

I now present the test statistics used in four of the most prominent

combined significance tests.

Stouffer's method: Adding z's. This method for summarizing p's involves adding the standard normal deviates or z scores associated with the k independent p values (Stouffer et al., 1949). The sum of the z scores is divided by the square root of k, the number of studies, which is the standard deviation of a sum of k standard normal deviates. The test statistic for this summary is

$$\sum_{i=1}^{k} z(p_i) \; / \; \sqrt{k}, \tag{6}$$

where $z(p_i) = \Phi^{-1}(p_i)$ and Φ is the cumulative distribution function of the standard normal distribution. Thus $z(p_i)$ represents the standard normal deviate associated with upper-tail probability p_i from the ith study. Under the null hypothesis (4) the statistic (6) has a standard normal distribution.

Fisher's method: Adding logs. Sir Ronald A. Fisher (1932) proposed the use of the logarithm of the product of the sample probabilities to represent the significance of an aggregation of independent results. The statistic used in this method is

$$-2 \ln \left(\prod_{i=1}^{k} p_i \right), \tag{7}$$

or

$$-2 \sum_{i=1}^{k} \ln (p_i).$$

When the null hypothesis in (4) is true, the statistic (7) is distributed as a chi-square random variable with 2k degrees of freedom, where k is the number of independent p values being summarized.

George's logit method. The logit method is one of the most recently suggested methods for summarizing p values (George, 1977). The method involves transforming each p_i value to a logistic variable (which will be symmetric about zero). George's statistic is

$$-K_1 \sum_{i=1}^{k} \log (p_i / 1 - p_i), \tag{8}$$

where $K_1 = \sqrt{[3(5k+4)/(k\pi^2(5k+2)]}$. When the null hypothesis in (4) is true the summary shown in (8) is approximately distributed as a t variable with $5k + 4$ degrees of freedom. The t approximation appears to be very accurate in small samples.

Testing the mean p. Edgington (1972b) developed a method for summarizing significance values based on the mean of the k independent p values. When the null hypothesis shown in (2) is true for the ith study, then p_i is a uniform random variable distributed on the interval from zero to one. Thus the population mean of a set of such uniform variables is 1/2, and the variance is 1/12.

Edgington's test statistic is

$$(.5 - \bar{p})\sqrt{12k}, \tag{9}$$

where $(.5 - \bar{p})$ represents the deviation of the mean p ($\bar{p}$) from its expected population value of 0.5, and $1/\sqrt{12k}$ is the standard error of the mean p. By the Central Limit Theorem the statistic (9) is approximately a standard normal variable when k is large and the null hypothesis in (4) is true. Thus an approximate test is obtained by comparing (9) to critical values from a table of standard normal deviates.

Power of Tests of Combined Significance and Asymptotic Distributions

There are many combined significance tests (as well as other statistical methods for summarizing study results), thus reviewers wishing

to use quantitative summaries must choose among the available tests. One way of selecting among available summaries is to evaluate their power to detect a plausible and interesting pattern of study results, choosing the test which is most likely to detect that pattern. Most of the familiar statistics used in single studies (e.g., t and F tests) are used because they are the most powerful at detecting a wide variety of (and sometimes all) alternative hypotheses. However, none of the combined significance tests is uniformly most powerful against all alternatives (Wallis, 1942).

The power of a statistical test depends on the values of the population parameters under the specified alternative hypothesis. When statistical tests for analyzing single population parameters are examined, the alternative hypothesis can be easily described via the size of the difference between the proposed null value and an alternative parameter value. Statistical tests concerning several population parameters (e.g., parameters from k studies) involve complicated alternative hypotheses because there are many possible deviations from the null condition. For example, a single parameter may disagree with its null value, or all k parameters may be nonnull. In the latter case, the k parameter values may or may not all be equal. This multiplicity of possible alternative hypotheses complicates the process of examining power for summaries of results from series of studies.

In order to understand the sensitivity of combined significance test procedures to various deviations from the null hypothesis, it would be desirable to obtain the sampling distribution of the combined test statistic when the null hypothesis is false. Unfortunately the noncentral sampling distributions of combined test statistics are unknown. Moreover, the exact noncentral distributions are likely to be extremely complex, if not intractable. Thus derivation of exact distributions for combined test

statistics is unlikely to be fruitful.

Because of the mathematical complexity of exact distributions for combined significance tests when the null hypothesis is false, alternative approaches to computing power have been sought. Statisticians have used proof, enumeration of all outcomes, numerical evaluation of power functions, and simulation to explore the power of tests of combined significance.

Another approach, which gives results of considerable generality, is the derivation of simple approximations to the noncentral distributions of combined test statistics based on their asymptotic distributions. The large-sample distributions derived in this paper are simple (relative to the exact ones) and are normal, thus making approximate power calculations straightforward.

Asymptotic Distributions

In this section the noncentral asymptotic distributions for four combined significance test criteria are derived. I first present a multivariate convergence theorem needed to derive the distributions. The proof is provided in Appendix B. Second, I state a result of Lambert (1978) which is used to derive the asymptotic distributions. In the third part of this section the distributions are derived, and last, the accuracy of the distributions is examined.

Convergence Theorem

The combined test statistics introduced above are continuous functions of vectors of k independent p values. A general procedure known as the δ method (Cramér, 1946) can often be used to obtain the asymptotic distribution of a continuous function of several random variables. However, when the function of interest involves both the random variables and the sample size N, the δ method cannot be used to derive the distribution for that function. Because some of the combined significance tests have that character, a modification of the δ method is needed. The following

convergence theorem is used to derive the limiting distributions for continuous functions of both the data and $\underline{N}$.

Theorem. Let G be a $\underline{k}$-dimensional statistic $(\underline{G}_1, \ldots, \underline{G}_{\underline{k}})'$, where $\underline{G}_{\underline{i}}$ is the $\underline{i}$th independent statistic and is based on a sample of size $\underline{n}_{\underline{i}}$. Define $\underline{N} = \Sigma\, \underline{n}_{\underline{i}}$ and let $\pi_{\underline{i}} = \underline{n}_{\underline{i}}/\underline{N}$, $\underline{i} = 1, \ldots, \underline{k}$, be fixed as $\underline{N} \to \infty$. Also define $\Gamma = (\gamma_1, \ldots, \gamma_{\underline{k}})'$ where $\gamma_{\underline{i}}$ is the expectation of $\underline{G}_{\underline{i}}$.

Let $\underline{N}^{\underline{a}}(G - \Gamma)$ where $\underline{a} > 0$ have an asymptotic distribution which is $\underline{k}$-variate normal with mean vector 0 and diagonal covariance matrix Σ, where $\Sigma = \mathrm{diag}(\sigma^2(\gamma_1), \ldots, \sigma^2(\gamma_{\underline{k}}))$. Formally this can be written

$$\underline{N}^{\underline{a}}(G - \Gamma) \xrightarrow{L} \underline{X} \sim N(0, \Sigma). \tag{10}$$

Further, let f be a real-valued function of the $\underline{k}$ statistics and $\underline{N}$, which is at least twice differentiable and has continuous first derivatives such that

$$\lim_{\underline{N} \to \infty} \frac{\partial f(\Gamma, \underline{N})}{\partial \underline{G}_{\underline{i}}} = \underline{F}_{\underline{i}}(\Gamma), \quad 0 \leq |\underline{F}_{\underline{i}}(\Gamma)| < \infty, \tag{11}$$

for $\underline{i} = 1, \ldots, \underline{k}$. The $\underline{k}$-dimensional column vector of these limits is denoted as F. The $\underline{k}$ elements of F must not all equal zero.

Then the asymptotic distribution of $\underline{N}^{\underline{a}}[f(G, \underline{N}) - f(\Gamma, \underline{N})]$ is given by

$$\underline{N}^{\underline{a}}[f(G, \underline{N}) - f(\Gamma, \underline{N})] \xrightarrow{L} \underline{Y} \sim N(0, \underline{V}), \tag{12}$$

where $\underline{V} = F'\, \Sigma\, F$, or equivalently $\underline{V} = \sum_{\underline{i}=1}^{\underline{k}} [\underline{F}_{\underline{i}}(\Gamma)\, \sigma(\gamma_{\underline{i}})]^2$.

A critical feature of the theorem is the requirement shown in (11), that the $\underline{k}$ first derivatives of the function of G and $\underline{N}$ have finite limits as $\underline{N}$ increases. Because these derivatives are directly involved in the

variance (V) of the function, this requirement is essential.

The Asymptotic Distribution of z(p)

Lambert (1978) stated without formal proof an expression for the sampling distribution of $z(p) = \Phi^{-1}(p)$, the one-tailed p value transformed by the inverse cumulative normal distribution function. Her result is as follows. Let T be the test statistic used to construct p, the one-tailed significance value, and suppose that T has an asymptotic normal distribution given by

$$\sqrt{n}\,(T - \theta) \overset{A}{\sim} N[0, \sigma^2(\theta)].$$

Then $z(p) = \Phi^{-1}(p)$ has an asymptotic normal distribution given by

$$\sqrt{n}\,[\, n^{-1/2}\, z(p) - \sqrt{2c(\theta)}\,] \overset{A}{\sim} N[0, \eta^2(\theta)]. \qquad (13)$$

where $c(\theta)$ is Bahadur's (1960) half-slope for the test T and

$$\eta^2(\theta) = [c'(\theta)\,\sigma(\theta)/2c(\theta)]^2. \qquad (14)$$

The parameter $\sigma^2(\theta)$ is the variance of the test statistic T and $\eta^2(\theta)$ is obtained from the asymptotic distribution of the logarithm of the p-value based on T (Lambert & Hall, 1982, 1983). When T represents a one-sample t test, then θ represents $\delta = \mu/\sigma$, $\sigma^2(\theta) = 1 + \delta^2/2$, the half-slope $c(\theta)$ is $[\log(1 + \delta^2)]/2$, and $\eta^2(\theta)$ equals $\delta^2(1 + \delta^2/2)/[(1 + \delta^2)\log(1 + \delta^2)]$.

The small-sample accuracy of this approximation to the distribution for z(p) was investigated computationally. For numerous combinations of sample sizes and effect sizes, approximate cumulative distribution function (cdf) values based on (13) were compared to numerical evaluations of the exact cdf of the untransformed p, obtained from the t statistic. (For details and more extensive results of the accuracy study see Becker (1985)).

Table 1 presents the accuracy results for p's from one-sample t tests with n = 10. Even for this small sample size the approximation is quite accurate, particularly for small values of δ. Furthermore, accuracy increases for all values of δ as n increases. The largest absolute difference in cdf's found in any case examined is only .10, for n = 10 and δ = 2.00. Since questions about the power of summaries of study results are most interesting when the effects (δ's) are small, the inaccuracy in this approximation will not influence its usefulness in most reasonable cases.

Insert Table 1 about here

Distributions for the Four Test Statistics

In this section the asymptotic distributions of test statistics associated with the four combined significance tests are derived. In most cases the actual distribution obtained is that of a one-to-one function of the test statistic, which can, in turn, be used to obtain the distribution of the desired test statistic.

To obtain finite first derivatives for the Fisher, logit, and mean-p summaries an additional constraint on the parameters was needed. Specifically, the noncentrality parameter for the ith study, $\sqrt{N}\ \delta_i$, was constrained to be fixed as N increases. This is the constraint used in deriving the usual asymptotics for the noncentral t distribution. The fixed parameter for the ith study is denoted as $\xi_i = \sqrt{N}\ \delta_i$.

Below I present for each method its test statistic, the function of that statistic used to obtain the distribution, and the relationship of the function to the original test statistic presented above. Then the mean and variance for each function are given.

Stouffer's method: Adding z's. This statistic has the simplest

asymptotic distribution because it is a linear combination of $z(p_i)$ values. The slight modification of the summary used to derive an approximate distribution is

$$T_S = \sum_{i=1}^{k} z(p_i)/\sqrt{Nk}, \tag{15}$$

which is simply the test shown in (6) divided by $\sqrt{N}$. The expectation of the function (15) is

$$\theta_S = \sum_{i=1}^{k} \sqrt{\pi_i \, 2 \, c(\delta_i)/k}. \tag{16}$$

Define $N = \Sigma n_i$ and let $\pi_i = n_i/N$, $i = 1, \ldots, k$. When $N \to \infty$ and the π_i, $i = 1, \ldots, k$, remain fixed, T_S has an asymptotic distribution given by

$$\sqrt{N}(T_S - \theta_S) \overset{A}{\sim} N(0, \sigma_S^2), \quad \text{where}$$

$$\sigma_S^2 = k^{-1} \sum_{i=1}^{k} \eta^2(\delta_i), \tag{17}$$

and $\eta^2(\delta_i)$ is given by (14).

Fisher's method: Adding logs. The summary statistic for Fisher's summary is the product of the p values (from formula (7)), and the statistic used to derive the asymptotic distribution is

$$T_F = -\phi^{-1} \left[\prod_{i=1}^{k} p_i \right]/\sqrt{N}, \tag{18}$$

which is related to the original Fisher summary (7), denoted here as F, via

$$-\phi^{-1} \left[\exp(-F/2)\right]/\sqrt{N}. \tag{19}$$

The expected value of the statistic (18) is

$$\theta_F = -\Phi^{-1}[\prod_{i=1}^{k} \Phi(-\sqrt{N \pi_i 2 c(\delta_i)})]/\sqrt{N}, \tag{20}$$

which can be easily seen to resemble equation (18); each p_i has been replaced by its expectation $\Phi(-\sqrt{N \pi_i 2 c(\delta_i)})$. Thus when $N \to \infty$ such that the π_i and $\xi_i = \sqrt{N}\,\delta_i$, $i = 1, \ldots, k$, remain fixed, then

$$\sqrt{N}(T_F - \theta_F) \overset{A}{\sim} N(0, \sigma_F^2), \quad \text{where}$$

$$\sigma_F^2 = \sum_{i=1}^{k} \eta^2(\delta_i) \frac{\exp\{-\pi_i \xi_i^2\} \left(\prod_{j=1}^{k} \Phi(-\sqrt{\pi_j}\,\xi_j)\right)^2}{\Phi(-\sqrt{\pi_i}\,\xi_i)^2 \exp\{-[\Phi^{-1}(\prod_{j=1}^{k} \Phi(-\sqrt{\pi_j}\,\xi_j))]^2\}}. \tag{21}$$

Logit method. The logit method uses a summary statistic involving the log of the product of $p/(1-p)$ terms. The function of the summary statistic used to derive the asymptotic distributions is

$$T_L = -K_1 [\log(\prod_{i=1}^{k} p_i) - \log(\prod_{i=1}^{k} (1 - p_i))]/\sqrt{N}, \tag{22}$$

where K_1 is defined as for (8) above. This function is simply the formula (8) divided by $\sqrt{N}$. The expectation of (22) is

$$\theta_L = -K_1 [\log(\prod_{i=1}^{k} \Phi(-\sqrt{N \pi_i 2c(\delta_i)})) - \log(\prod_{i=1}^{k} (1-\Phi(-\sqrt{N \pi_i 2c(\delta_i)})))]/\sqrt{N}. \tag{23}$$

When $N \to \infty$ and π_i and $\xi_i = \sqrt{N}\,\delta_i$, $i = 1, \ldots, k$, remain fixed,

$$\sqrt{N}\,(T_L - \theta_L) \overset{A}{\sim} N(0, \sigma_L^2), \quad \text{where}$$

$$\sigma_L^2 = \underline{K}_1^2 \sum_{\underline{i}=1}^{\underline{k}} \eta^2(\delta_{\underline{i}}) \frac{\exp(-\pi_{\underline{i}} \xi_{\underline{i}}^2)}{2\pi\, \Phi(-\sqrt{\pi_{\underline{i}}}\, \xi_{\underline{i}})^2\, [1 - \Phi(-\sqrt{\pi_{\underline{i}}}\, \xi_{\underline{i}})]^2}. \tag{24}$$

Testing the mean p. The summary statistic for Edgington's (1972b) mean-p summary method, as well as for another method he suggested (1972a), is the mean of the p values. The function used to derive an approximate distribution of the mean-p summary is

$$T_M = -\Phi^{-1}(\bar{p})/\sqrt{N}, \tag{25}$$

which relates to the original summary via

$$-\Phi^{-1}(0.5 - M/\sqrt{12k})/\sqrt{N}, \tag{26}$$

where M denotes the expression given in (9). The expectation of the function (25) is

$$\theta_M = -\Phi^{-1}[k^{-1} \sum_{i=1}^{k} \Phi(-\sqrt{N\, \pi_i\, 2\, c(\delta_i)})]/\sqrt{N}. \tag{27}$$

Again the similarities between the random variable (25) and its expectation (27) can be easily seen. When $N \to \infty$ such that the π_i and $\xi_i = \sqrt{N}\, \delta_i$, for $i = 1,\ldots,k$, remain fixed, then T_M has an asymptotic distribution given by

$$\sqrt{N}\, (T_M - \theta_M) \overset{A}{\sim} N(0, \sigma_M^2), \text{ where}$$

$$\sigma_M^2 = k^{-2} \sum_{i=1}^{k} \eta^2(\delta_i) \exp\{-\pi_i \xi_i^2 + [\Phi^{-1}(k^{-1} \sum_{j=1}^{k} \Phi(-\sqrt{\pi_j}\, \xi_j))]^2\}. \tag{28}$$

Other Combined Significance Tests

It is useful to note that there are other combined significance tests for which asymptotic distributions can be derived, and some for which

distributions are impossible to obtain using the illustrated methods. For example, distributions for weighted versions of the Stouffer and logit methods follow directly from the above results.

On the other hand, Tippett's (1931) test based on the minimum p is a discrete function of the vector of p's, and thus does not conform to the requirements of this convergence theorem. For other summaries it may be impossible to obtain finite limits for the first derivatives of any functions of the test statistic. However, as a general approach, the convergence theorem presented above can be widely applied.

Simulation Study of the Distributions

Although large-sample approximations such as those given above may be reasonably accurate in small samples, it is unwise to use them without evaluating their accuracy. One way to evaluate the accuracy of a large-sample approximation is via a simulation study. Consequently a simulation study was conducted to determine the accuracy of the large-sample theory given in this paper.

The study was based on various combinations of N, π's, and δ values for $k = 2$ and $k = 5$ sets of studies. Values of N and π_1 through π_k were selected on the basis of evidence from an empirical study of sample sizes in research reviews (Becker, 1985). For each combination of N, π's, and δ's, 5000 sets of either two or five t variates were generated and their p values were obtained via International Mathematical and Statistical Libraries (1982) subroutines GGNML, GGCHS, and MDTD.

These p's were used to calculate the functions representing the four summaries (i.e., formulas (15), (18), (22), and (25), hereafter referred to as the "summary functions"), which were then standardized with the appropriate means and standard deviations derived from their asymptotic distributions. If the asymptotic distribution is accurate, these "standardized" variates should be approximately normally distributed with a

mean of zero and variance of one. Also the empirical proportions of the standardized summary-function values which exceed any given value should correspond to proportions exceeding that value in the standard normal distribution.

Initial computations showed that the asymptotic distributions presented above for the Fisher, mean-$\underline{p}$, and logit tests often differed systematically from the standard normal distribution in finite samples. The means and variances of the statistics were modified, which improved the accuracy of the distributions, though the distribution based on the mean $\underline{p}$ was still quite inaccurate.

Adjustments to the means and variances were derived from graphs of the deviations of the empirical moments for each function (from the initial simulations) from 0 and 1, respectively. Functions based on $\underline{n}$ and δ were found which reproduced fairly closely the empirical variances of the mean $\underline{p}$ and logit summaries and the errors in the terms of the means for all four tests. Table 2 presents these functions.

Insert Table 2 about here

A simulation based on random numbers different from those used to derive the modifications showed that the means and variances of the modified standardized summaries were much closer to zero and one, respectively, than those of the unmodified standardized summaries.

A second aspect of the behavior of the tests which was studied was the proportion of times that simulated random values of the summary functions, (appropriately standardized) were larger than percentiles of the standard normal distribution. Forty-five different configurations of $\underline{N}$, π's, and δ's were studied, and for each configuration, 5000 replications of $\underline{k} = 2$ $\underline{p}$

values were generated. For each configuration, the proportions of simulated summary function values that exceeded normal percentiles (the "empirical exceedance proportions") were noted. Six normal percentiles were observed, the values of p^* were .75, .50, .25, .10, .05, and .01, where $1 - p^*$ is the percentile rank of the normal percentile. If the distribution of a summary is relatively normal, the empirical exceedance proportions should be very close to p^*.

Table 3 shows the percentages of empirical exceedance proportions differing from p^* for the four tests at six nominal p^* values. The number of cases represents the number of configurations of N, π's, and δ's summarized in the table; each "case" represents a set of empirical exceedance proportions based on 5000 replications. Fewer exceedance proportions differed significantly from the nominal p^* for the modified tests than for the unmodified tests, though the performance of the distribution of the statistic based on the mean p was still quite poor in some cases.

Insert Table 3 about here

The distribution for the test statistic proposed by Stouffer shows the most agreement between nominal and empirical proportions, as would be expected. The distribution of the summary based on the mean p is least normal, as it was prior to modification. It shows much disagreement between nominal and empirical exceedance proportions when sample sizes and effect sizes are unequal. The modified Fisher summary seems to show rather good agreement in the tails (i.e., for small p^* values) and for cases in which tne two effect sizes are equal. The modified logit summary, on the other hand, appears most nonnormal in the tails, with a minimum of 48 percent of

the exceedance proportions differing from $\underline{p}^*$ when $\underline{p}^* \leq .05$.

The modifications were also studied in several simulations for the case in which $\underline{k} = 5$ studies were summarized. Though the modifications did not improve the distributions as much as in the $\underline{k} = 2$ case, the modified distributions for $\underline{k} = 5$ compared favorably with the unmodified distributions.

A final component of the simulation study was a comparison of simulated size and power values (which should represent fairly well the "true" α and power of each test) to values estimated using the (modified) distributions given above. Table 4 presents estimated and simulated size values at three nominal α^* levels for several sample size configurations. (In all cases the δ values were equal to 0.00001 to represent the null hypothesis condition. Using δ values of exactly zero produced computational difficulties with $\underline{c}(\delta)$ and $\eta(\delta)$, thus a very small δ was substituted.) Note that these size values should be equal to or very near α^*.

Insert Table 4 about here

For these selected α^* levels the distributions are fairly accurate under the null hypothesis; the simulated ("true") size values are quite close to the nominal α^* values for the four tests. The Stouffer method produces estimated sizes that are identically equal to α^*. The estimates from the Fisher-test distribution are consistently larger than α^* for $\underline{k} = 10$, though for $\underline{k} = 2$ and $\underline{k} = 5$ they are quite close. The logit estimates and simulated size values both appear quite close to α^* in most cases shown. The estimates of size for the mean-$\underline{p}$ test for these α^* levels appear to be in general too large, relative to both α^* and the simulated values.

Insert Table 5 about here

Maximum differences between estimated and simulated power values indicate the extent to which the power functions based on the approximate distributions give power results which will deviate from their "true" values. Table 5 presents values of these maximum differences across six nominal α^* levels (α^* is .75, .50, .25, .10, .05, or .01). Two models for study outcomes are examined; the one reported above in which all δ values equal 10^{-5}, and a second in which all δ values equal 0.25.

For pairs of study results the maximum difference between the estimated and simulated values in the cases shown is .03. The differences are larger for $\underline{k} = 5$ and again the distribution of the mean $\underline{p}$ statistic appears least accurate, with discrepancies of more than .10 between the estimated and simulated values in some cases. Becker (1985) gives further details of the simulation study.

Power Computation Examples

The asymptotic distributions presented above may be applied in a comprehensive study of power across many alternatives. Such a study can provide generalizations that may enable researchers to select a combined significance test on the basis of power differences for situations similar to one of interest. Becker (1985) presented such results.

Another use of the distributions is to make specific computations of power of the tests to detect plausible effects in a set of particular studies. Thus a researcher contemplating the use of a combined significance test in a particular research review can estimate the power of each test to detect an interesting outcome, and can then choose the most powerful test.

In this section the distributions are used to calculate power to detect several plausible outcomes for a set of studies from Table F in the Appendix of Gage's (1978) book The scientific basis of the art of teaching. The five studies in this data set examined the relationship between teacher criticism

and disapproval and student vocabulary achievement.

Vocabulary and Teacher Criticism and Disapproval

Table 6 presents the authors, sample sizes (n_i), empirical sampling fractions ($\hat{\pi}_i$), sample correlations (r_i) and corresponding effect size estimates (d_i) for the five studies from Gage (1978). The d_i estimates were obtained from the correlations using the relationship

$$d = -2r / \sqrt{1 - r^2}.$$

This relationship is based on a very accurate approximation of the noncentral t distribution by this function of the correlation coefficient (see Kraemer, 1983). Signs of the sample estimates were reversed here so that the majority of the d_i values would be positive. The total sample size (N) for this set of $k = 5$ studies is 217.

Insert Table 6 about here

The five studies examined the relationship between observational measures of teacher criticism and achievement in vocabulary as measured by standardized tests such as the Stanford Achievement Test and the Metropolitan Achievement Tests. The research hypothesis tested in each one was that there would be a negative relationship between criticism or disapproval and vocabulary achievement. More teacher criticism was thought to be associated with poorer student achievement in vocabulary. Thus the statistical hypotheses in the ith study were

$$H_0: \rho_i = 0 \text{ and } H_1: \rho_i < 0,$$

or in terms of the related effect size δ_i,

$$H_0: \delta_i = 0 \text{ and } H_1: \delta_i > 0.$$

The null hypothesis for the series of five studies (in terms of δ) is

$$H_0: \ \delta_1 = \ldots = \delta_5 = 0,$$

and the alternative is that at least one δ_i value is positive (i.e., that at least one ρ_i is negative).

One plausible outcome is that all of the studies would share a population correlation equal to the size of the "typical" obtained correlation. (Note that reviewers should not examine study outcomes to develop a hypothetical parameter value for use in power computations, but should use an *a priori* theory or prior research to justify such a value. We do so here only for the purpose of illustrating the calculations.) The median r value is -.16, which corresponds to an effect size estimate of d = 0.32. We can use the approximate distributions to determine how likely each test would be to detect the outcome in which all five studies have a population correlation of ρ_i = -.16 (or δ_i = 0.32). A test at the α = .05 significance level of the null model given above is considered as an illustration.

For each test, the appropriate null-hypothesis critical value is obtained and then transformed into the metric of the summary function. For example, for the Fisher test the .05 value of the chi-square distribution with $2k$ = 10 degrees of freedom is 18.31, which when transformed into the metric of the function used to derive the distribution (i.e., transformed via equation (18)) equals 0.252. The four transformed critical values values are standardized by their appropriate means and standard deviations to give the z values shown in the first column of Table 7.

Insert Table 7 about here

The power estimates are then obtained from the standard normal table. The power values associated with these $\underline{z}$'s are all quite high, indicating that all of the significance tests are almost certain to detect the hypothesized outcome that $\rho_1 = \ldots = \rho_5 = -.16$.

We examine another alternative hypothesis in which all studies again share a common correlation, but where the common ρ is -.08, or half the size of the previously proposed result. Carrying out similar calculations to those detailed above we obtain the $\underline{z}$ scores shown in the second set of columns in Table 7.

For this alternative the power values are smaller, and somewhat more distinct than in the first example. The Stouffer method has the highest power estimate, with a 75.8 percent chance of detecting the hypothesized outcome. The Fisher test has only slightly less power, with a nearly 74 percent chance of rejecting H_0, followed by the logit test with a power estimate of 71 percent. The mean-$\underline{p}$ test, however, appears much less likely to reject H_0 with an estimated power of 46 percent. Though results of the simulation of power indicate that power estimates for the mean $\underline{p}$ test are generally smaller than they should be, the differences between estimated and simulated values are seldom large enough (i.e., greater than .26) for the mean $\underline{p}$'s actual power to be close to that of the other tests.

In a final example the studies are grouped into two sets, those expected (presumably on the basis of some theory) to show a stronger relationship between teacher criticism and achievement and those expected to show a weaker relationship. Imagine that the three larger studies are expected to have population correlations of only -.05 (slightly less than their actual mean correlation) and the smaller studies are expected to show ρ values of -.16 (as in the first example).

The last columns of Table 7 show that the Fisher test is most likely to detect this alternative, with a power estimate of 86 percent. The Stouffer

and logit tests are not much less powerful, but again the mean-$\underline{p}$ method's power estimate is quite unsatisfactory and low.

These few examples illustrate the application of the asymptotic distributions in the calculation of power for combined significance tests. By substituting the appropriate parameters in place of δ, $\underline{c}(\delta)$, and $\eta^2(\delta)$, power estimates can be obtained for summaries involving $\underline{p}$ values from many other statistical tests. Appendix B gives formulas for the parameters which would be used in computing power for summaries of $\underline{p}$'s from two-sample $\underline{t}$ tests.

Summary

Exact distributions for sample significance values and functions of those values are mathematically complex. Thus, obtaining exact power values for combined significance tests has until now required either numerical computation or limitation of the set of outcomes to those for which power computations are possible through enumeration. Approximate power values have been available through the use of extensive simulation studies, though those have depicted power for a limited number of configurations of study features (e.g., sample sizes, numbers of studies, potential outcomes, etc.).

The provision of asymptotic distributions for the tests of combined significance allows researchers to estimate power values for specific situations or study configurations. These values in turn can enable researchers to choose a combined significance test on the basis of its ability to detect a particular pattern of outcomes for studies like those being reviewed (i.e., in terms of size and test statistic used). Many different patterns of population effect-sizes can be associated with significant results from the combined significance tests. These distributions should enable researchers to select a test that will best detect interesting and useful outcomes.

References

Bahadur, R. R. (1960). On the asymptotic efficiency of tests and estimates. Sankhyá, 22, 229-252.

Becker, B. J. (1985). Applying tests of combined significance: Hypotheses and power considerations. Unpublished doctoral dissertation, University of Chicago.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. Journal of Personality and Social Psychology, 37, 131-146.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. Psychological Bulletin, 87, 442-449.

Cramér, H. (1946). Mathematical methods of statistics. Princeton, NJ: Princeton University Press.

Edgington, E. S. (1972a). An additive method for combining probability values from independent experiments. Journal of Psychology, 80, 351-363.

Edgington, E. S. (1972b). A normal curve method for combining probability values from independent experiments. Journal of Psychology, 82, 85-89.

Fisher, R. A. (1932). Statistical methods for research workers. (4th ed.). London: Oliver and Boyd.

Gage, N. L. (1978). The scientific basis of the art of teaching. New York: Teachers College Press.

George, E. O. (1977). Combining independent one-sided and two-sided statistical tests: Some theory and applications. Doctoral dissertation, University of Rochester.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Harris, A. J., & Serwer, B. L. (1966). Comparison of reading approaches in first grade teaching with disadvantaged children (the CRAFT Project) (USOE Cooperative Research Project No. 2677). New York: City University of New York.

Harris, A. J., Morrison, C., Serwer, B. L., & Gold, L. (1968). A continuation of the CRAFT Project: Comparing reading approaches with disadvantaged urban Negro children in primary grades (USOE Project No. 5-0570-2-12-1). New York: Division of Teacher Education, City University of New York.

International Mathematical and Statistical Libraries, Inc. (1982). IMSL Library 1. (8th ed.). Houston, TX: Author.

Jones, L. V. & Fiske, D. W. (1953). Models for testing the significance of combined results. Psychological Bulletin, 50, 375-382.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. Journal of Educational Statistics, 8, 93-101.

Lambert, D. M. (1978). P values: Asymptotics and robustness. Doctoral dissertation, University of Rochester.

Lambert, D. M., & Hall, W. J. (1982). Asymptotic lognormality of p values. The Annals of Statistics, 10, 44-64.

Lambert, D. M., & Hall, W. J. (1983). Correction to "Asymptotic lognormality of p values." The Annals of Statistics, 11, 348.

Mosteller, F., & Busn, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), Handbook of social psychology: Vol. 1. Theory and method. Cambridge, MA: Addison-Wesley.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Soar, R. S. (1966). An integrative approach to classroom learning (Final Report, Public Health Service Grant No. 5-R11-MH 01096 & National Institute of Mental Health Grant No. 7-R11-MH02045). Philadelphia: Temple University.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). The American soldier: Adjustment during army life. (vol. 1). Princeton: Princeton University Press.

Tippett, L. H. C. (1931). The methods of statistics. (1st ed.). London: Williams & Norgate.

Wallen, N. E. (1966). Relationships between teacher characteristics and student behavior, Part 3 (USOE Cooperative Research Project No. SAE OE5-10-181). Salt Lake City: University of Utah.

Wallis, W. A. (1942). Compounding probabilities from independent significance tests. Econometrica, 10, 229-248.

White, K. R. (1982). The relationship of socioeconomic status and academic achievement. Psychological Bulletin, 91, 461-481.

Table 1

Accuracy of the Inverse Normal Approximation to the cdf of p

n	δ	Maximum absolute difference	Mean absolute difference	Mean difference
10	0.10	.00	.00	.00
	0.25	.01	.00	.00
	0.30	.01	.00	.00
	0.40	.01	.01	.01
	0.50	.01	.01	.01
	0.75	.02	.01	.01
	1.00	.03	.02	.02
	1.25	.04	.02	.02
	1.50	.06	.03	.03
	1.75	.08	.03	.03
	2.00	.10	.04	.04

Table 2

Modifications to Distributions for Combined Significance Tests

Method and conditions	Value added to ith term of mean	Value by which ith variance term is multiplied
Stouffer		
$n_i \leq 30$ and $\delta_i \geq 0.5$	-0.025	1.0
Otherwise	0.000	1.0
Fisner	$0.11 - 0.143\delta_i$	1.0
Logit[1]		
$n_i < 50$	$K_1(0.12 - 0.75(\delta_i - 0.40)^2)$	$1.35 - 0.68\delta_i$
$n_i \geq 50$	$K_1(0.15 - 0.60(\delta_i - 0.50)^2)$	$1.35 - 0.68\delta_i$
Mean p		
$n_i \leq 25$,		
$\delta_i \leq 0.5$	$-0.12 + 0.48(\delta_i - 0.50)^2$	$0.70 + 0.60\delta_i + \delta_i^2 - 1.25\delta_i^3$
$\delta_i > 0.5$	-0.14	$0.70 + 0.60\delta_i + \delta_i^2 - 1.25\delta_i^3$
$n_i > 25$,		
$\delta_i \leq 0.5$	$-0.17 + 0.68(\delta_i - 0.50)^2$	$0.75 + 0.75\delta_i + 0.50\delta_i^2 - 0.75\delta_i^3$
$\delta_i > 0.5$	-0.18	$0.75 + 0.75\delta_i + 0.50\delta_i^2 - 0.75\delta_i^3$

[1] K_1 is defined as for (8).

Table 3

Percentages of Exceedance Proportions Significantly Different From p* For Modified Summaries

Sampling fractions	Effect sizes	Number of cases	Nominal probability level (p*) .75	.50	.25	.10	.05	.01
Stouffer method								
$\pi_1 = \pi_2$	$\delta_1 = \delta_2$	31	0	6	3	3	0	0
$\pi_1 = \pi_2$	$\delta_1 \neq \delta_2$	4	0	0	0	0	0	0
$\pi_1 \neq \pi_2$	$\delta_1 = \delta_2$	3	0	0	0	0	0	0
$\pi_1 \neq \pi_2$	$\delta_1 \neq \delta_2$	7	0	0	0	0	0	0
Fisher method								
$\pi_1 = \pi_2$	$\delta_1 = \delta_2$	31	16	6	6	6	3	3
$\pi_1 = \pi_2$	$\delta_1 \neq \delta_2$	4	50	25	50	25	50	0
$\pi_1 \neq \pi_2$	$\delta_1 = \delta_2$	3	0	0	0	0	0	0
$\pi_1 \neq \pi_2$	$\delta_1 \neq \delta_2$	7	29	29	29	14	14	0
Logit method								
$\pi_1 = \pi_2$	$\delta_1 = \delta_2$	31	45	0	16	29	48	68
$\pi_1 = \pi_2$	$\delta_1 \neq \delta_2$	4	45	0	16	29	48	68
$\pi_1 \neq \pi_2$	$\delta_1 = \delta_2$	3	0	0	0	0	67	100
$\pi_1 \neq \pi_2$	$\delta_1 \neq \delta_2$	7	71	0	0	43	71	86
Mean p method								
$\pi_1 = \pi_2$	$\delta_1 = \delta_2$	31	45	35	23	0	0	6
$\pi_1 = \pi_2$	$\delta_1 \neq \delta_2$	4	100	100	100	75	100	100
$\pi_1 \neq \pi_2$	$\delta_1 = \delta_2$	3	67	67	33	33	0	33
$\pi_1 \neq \pi_2$	$\delta_1 \neq \delta_2$	7	100	100	86	100	86	86

Note: Simultaneous tests at the .01 significance level are used.

Table 4

Estimated and Simulated Size Values

Sample sizes	$\alpha^* = .10$ Estimated size	$\alpha^* = .10$ Simulated size	$\alpha^* = .05$ Estimated size	$\alpha^* = .05$ Simulated size	$\alpha^* = .01$ Estimated size	$\alpha^* = .01$ Simulated size
Stouffer method						
Case 1:	.100	.104	.050	.051	.010	.009
Case 2:	.100	.100	.050	.049	.010	.011
Case 3:	.100	.102	.050	.050	.010	.011
Case 4:	.100	.097	.050	.051	.010	.011
Case 5:	.100	.101	.050	.052	.010	.010
Case 6:	.100	.098	.050	.046	.010	.009
Fisher method						
Case 1:	.098	.103	.047	.052	.009	.010
Case 2;	.098	.102	.047	.050	.009	.010
Case 3:	.106	.104	.051	.049	.009	.009
Case 4:	.106	.101	.051	.055	.009	.011
Case 5:	.198	.102	.108	.052	.024	.011
Case 6:	.198	.092	.108	.046	.024	.008

Table 4 (continued)

Estimated and Simulated Size Values

	$\alpha^* = .10$		$\alpha^* = .05$		$\alpha^* = .01$	
Sample sizes	Estimated size	Simulated size	Estimated size	Simulated size	Estimated size	Simulated size
Logit method						
Case 1:	.112	.103	.055	.053	.009	.010
Case 2:	.112	.102	.055	.049	.009	.011
Case 3:	.108	.100	.054	.049	.010	.010
Case 4:	.108	.098	.054	.050	.010	.011
Case 5:	.107	.101	.054	.054	.011	.010
Case 6:	.107	.095	.054	.048	.011	.008
Mean $\underline{p}$ method						
Case 1:	.115	.119	.049	.053	.000	.002
Case 2:	.119	.111	.052	.051	.001	.001
Case 3:	.135	.102	.074	.050	.015	.009
Case 4:	.132	.100	.071	.052	.014	.009
Case 5:	.130	.105	.072	.052	.017	.010
Case 6:	.139	.102	.080	.051	.020	.009

Case 1: $\underline{n}_1 = \underline{n}_2 = 25$

Case 2: $\underline{n}_1 = 25, \underline{n}_2 = 75$

Case 3: $\underline{n}_1 = \ldots = \underline{n}_5 = 40$

Case 4: $\underline{n}_1 = 20, \underline{n}_2 = 20, \underline{n}_3 = 40, \underline{n}_4 = 40, \underline{n}_5 = 80$

Case 5: $\underline{n}_1 = \ldots = \underline{n}_{10} = 25$

Case 6: $\underline{n}_1 = \ldots = \underline{n}_{10} = 50$

Table 5

Maximum Differences Between Estimated and Simulated Power Values for Modified Summaries

	Sample	Effect	Method			
k	size	size	Stouffer	Fisher	Logit	Mean *p*
2	25	0.00001	.014	.011	.014	.014
2	25	0.25	.020	.024	.025	.028
2	50	0.00001	.004	.005	.021	.021
2	50	0.25	.006	.016	.044	.008
5	40	0.00001	.007	.036	.008	.003
5	40	0.25	.006	.035	.032	.165
5	60	0.00001	.006	.034	.007	.028
5	60	0.25	.004	.043	.035	.123

Note: All studies share one effect size, i.e., $\delta_i = \delta$, $i = 1,\ldots,k$.

Table 6

Studies From Gage (1978) Appendix Table F

Study	Sample size (n_i)	Sampling fraction ($\hat{\pi}_i$)	Correlation (r_i)	Effect size (d_i)
Harris & Serwer (1966)	48	.22	.06	-0.12
Harris et al. (1968)	38	.18	-.26	0.54
Soar (1966)	55	.25	-.16	0.32
Wallen (1966)	36	.17	-.38	0.82
Wallen (1966)	40	.18	-.14	0.29

Table 7

Power Results for Data from Gage (1978)

Method	$\delta = 0.32$		$\delta = 0.16$		$\delta_1 = 0.10, \delta_2 = 0.32$	
	$\underline{z}$	Power	$\underline{z}$	Power	$\underline{z}$	Power
Stouffer	-2.98	.999	-0.70	.758	-1.00	.841
Fisner	-2.75	.997	-0.63	.735	-1.08	.861
Logit	-2.36	.991	-0.55	.709	-0.87	.807
Mean $\underline{p}$	-1.63	.948	-0.05	.522	0.10	.460

Appendix A

In this appendix I present notation for the results of studies using two-sample comparisons, and outline how the distributions presented above would be applied to such results.

Consider a series of k studies each examining a treatment effect in a two-group design. Let Y^E_{ij} and Y^C_{ij} be the jth scores on some outcome variable Y in the experimental and control groups, respectively, of the ith study. Denote the ith experimental and control group sample sizes as n^E_i and n^C_i, and assume that in study i, Y^E_{ij} and Y^C_{ij} are normally distributed with means μ^E_i and μ^C_i and with common variance σ^2_i. Thus

$$Y^E_{ij} \sim N(\mu^E_i, \sigma^2_i), \quad \text{for } j = 1,\ldots,n^E_i, \; i = 1,\ldots,k,$$

and

$$Y^C_{ij} \sim N(\mu^C_i, \sigma^2_i), \quad \text{for } j = 1,\ldots,n^C_i, \; i = 1,\ldots,k.$$

In 1976 Glass proposed the use of the effect size or standardized mean difference as a scale-free measure of the difference between the experimental and control groups in a study. The effect size for the ith study is

$$g_i = \frac{\bar{Y}^E_i - \bar{Y}^C_i}{S_i}, \tag{29}$$

where $\bar{Y}^E_i$ and $\bar{Y}^C_i$ are the ith experimental and control group means, respectively, and S^2_i is the pooled within-groups estimate of σ^2_i. (Though Glass proposed using the control group standard deviation as the denominator of g_i, Hedges (1981) noted that when the assumption of equal population variances is satisfied the pooled standard deviation is more precise.)

Hedges (1981) noted that each g_i estimates a population effect size, δ_i, which may be written as

$$\delta_i = \frac{\mu^E_i - \mu^C_i}{\sigma_i}. \tag{30}$$

As above, μ_i^E and μ_i^C are the population means on Y for the ith experimental and control groups, respectively, and σ_i is the common population standard deviation of the Y scores in study i. The k parameters $\delta_1, \ldots, \delta_k$ are the population values about which inferences are made when sample effect sizes are analyzed.

The Significance Value

The relationship between the effect size and the sample significance value (or p value) arises through the relation of both to the two-sample t test. The effect size g is a simple linear transformation of the two-sample t test for the experimental versus control group comparison. The relationsnip of the effect size g to the test statistic t is shown as

$$t_i = \frac{\bar{Y}_i^E - \bar{Y}_i^C}{S_i / \sqrt{\tilde{n}_i}} = g_i \sqrt{\tilde{n}_i}, \qquad (31)$$

where $\tilde{n}_i = n_i^E \cdot n_i^C / (n_i^E + n_i^C)$. This t statistic has $n_i^E + n_i^C - 2$ degrees of freedom.

The sample significance value for the ith study is formally defined as

$$p_i = \int_{g_i \sqrt{\tilde{n}_i}}^{\infty} f(t, m_i)\, dt,$$

where $g_i \sqrt{\tilde{n}_i}$ is given in (31) and $f(t, m_i)$ is the probability density function for the central t distribution with $m_i = n_i^E + n_i^C - 2$ degrees of freedom.

Applying the Asymptotic Distributions

The asymptotic distributions for the tests of combined significance for combinations of studies using two-sample comparisons parallel those shown in the text. Only the formulas for $c(\delta)$ and $\eta(\delta)$ will differ. Formulas for $c(\delta)$ and $\eta(\delta)$ are given here for studies using two-sample t tests.

The formulas for $c(\delta)$ and $\eta(\delta)$

Lambert and Hall (1982, 1983) provide the formulas for $c(\delta)$ and $\tau(\delta)$ for the two-sample t test, from which the formula for $\eta(\delta)$ is derived. The similarities to the one-sample formulas are readily seen.

We first define n_{1i} and n_{2i} as the within-study sample sizes for the ith study. Let $n_i = n_{1i} + n_{2i}$, $\pi_{1i} = n_{1i}/n_i$, $\pi_{2i} = n_{2i}/n_i$, and let π_{1i} and π_{2i} be fixed as $n_i \to \infty$. Also define $\tilde{\pi}_i = \pi_{1i} \cdot \pi_{2i}$, the product of the sampling fractions for study i. Lambert and Hall give the slope $c(\delta_i)$ of the two-sample t test from study i as

$$c(\delta_i) = {}^1/_2 \ln (1 + \tilde{\pi}_i \delta_i^2), \tag{32}$$

and the variance of the logarithmically transformed p value from the t test as

$$\tau^2(\delta_i) = \frac{(1 + \frac{1}{2} \tilde{\pi}_i^2 \delta_i^2)\, \tilde{\pi}_i \delta_i^2}{(1 + \tilde{\pi}_i \delta_i^2)^2}.$$

Also, $\eta^2(\delta_i)$ is defined as $\tau^2(\delta_i)/2c(\delta_i)$, which by substitution gives us

$$\eta^2(\delta_i) = \frac{(1 + (1/2)\tilde{\pi}_i^2 \delta_i^2)\, \tilde{\pi}_i \delta_i^2}{(1 + \tilde{\pi}_i \delta_i^2)^2 \ln(1 + \tilde{\pi}_i \delta_i^2)}. \tag{33}$$

The values given in (32) and (33) are used in the formulas for the means and variances given in the text for the combined significance tests.

Appendix B

This appendix provides a proof of the convergence theorem stated in the text. In order to conserve space the theorem is not restated here.

Proof. Define $\nabla f(\Gamma, \underline{N})$ as the $\underline{k}$ by 1 vector of first derivatives of f (evaluated at Γ), and write the first-degree Taylor expansion of $f(G, \underline{N})$ about $G = \Gamma$ as

$$f(G, \underline{N}) = f(\Gamma, \underline{N}) + (G - \Gamma)'\nabla f(\Gamma, N) + \varepsilon \|G - \Gamma\|.$$

Equivalently this may be written as

$$f(G, \underline{N}) - f(\Gamma, \underline{N}) = (G - \Gamma)'\nabla f(\Gamma, \underline{N}) + \varepsilon \|G - \Gamma\|. \tag{34}$$

By Taylor's Theorem

$$\lim_{G \to \Gamma} \varepsilon = \frac{f(G, \underline{N}) - f(\Gamma, \underline{N}) - (G - \Gamma)'\nabla f(\Gamma, \underline{N})}{\|G - \Gamma\|} = 0.$$

Because ε can be expressed as a continuous function of $\|G - \Gamma\|$, for a given e there exists a corresponding δ_e such that $\|G - \Gamma\| < \delta_e$ implies that $|\varepsilon| < e$. Furthermore if $\|G - \Gamma\| < \delta_e$ implies that $|\varepsilon| < e$ then $P(\|G - \Gamma\| < \delta_e) \leq P(|\varepsilon| < e)$.

The statistic G is a consistent estimate for Γ because $E(G)$ equals Γ and the variance of G tends to zero as $\underline{N} \to \infty$. Thus $G \xrightarrow{P} \Gamma$ (i.e., G tends to Γ in probability) implies that the limit of $P(\|G - \Gamma\| > \delta)$ is zero as $\underline{N} \to \infty$ and for all δ. Thus

$$\lim_{\underline{N} \to \infty} P(|\varepsilon| > e) \leq \lim_{\underline{N} \to \infty} P(\|G - \Gamma\| > \delta_e) = 0,$$

or $\varepsilon \xrightarrow{P} 0$.

Formula (34) can now be rewritten as

$$\underline{N}^{a}[f(G, \underline{N}) - f(\Gamma, \underline{N})] - \underline{N}^{a}[(G - \Gamma)'\nabla f(\Gamma, \underline{N})] = \underline{N}^{a}\varepsilon\|G - \Gamma\|. \quad (35)$$

Then $\underline{N}^{a}\|G - \Gamma\| = (\sum [\underline{N}^{a}(\underline{G}_{i} - \gamma_{i})]^{2})^{1/2}$ which has a limiting distribution by assumption. Thus $\underline{N}^{a} \varepsilon \|G - \Gamma\| \xrightarrow{P} 0$, which implies that the left side of (35) also approaches zero in probability.

By assumption $\underline{N}^{a}(G - \Gamma)$ has a limiting distribution (that of X) and the elements of $\nabla f(\Gamma, \underline{N})$ approach the limits contained in the vector F as $\underline{N}$ increases. Thus the product $\underline{N}^{a}(G - \Gamma)'\nabla f(\Gamma, \underline{N})$, which is the second term in (34), also has a limiting distribution. The distribution is that of $X'F = \underline{Y}$, which is normal with mean zero and variance $\underline{V}$ given in (12) in the text.

Because $\underline{N}^{a}[f(G, \underline{N}) - f(\Gamma, \underline{N})] - \underline{N}^{a}[(G - \Gamma)'\nabla f(\Gamma, \underline{N})] \xrightarrow{P} 0$, and because the second term has a limiting distribution $\underline{N}^{a}[f(G, \underline{N}) - f(\Gamma, \underline{N})]$ also has a limiting distribution. It is the same as that for $\underline{N}^{a}[(G - \Gamma)'\nabla f(\Gamma, \underline{N})]$ or $X'F = \underline{Y}$. Thus

$$\underline{N}^{a}[f(G, \underline{N}) - f(\Gamma, \underline{N})] \xrightarrow{L} \underline{Y} \sim N(0, \underline{V}),$$

where $\underline{V}$ is defined in (12) in the text. ∎